DOCUMENT RESUME

ED 104 958 TM 004 418

AUTHOR Dryden, Russell E.; Frisbie, David A.
TITLE Comparative Reliabilities and Validities of Multiple Choice and Complex Multiple Choice Nursing Education Tests.
PUB DATE Apr 75
NOTE 11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D.C., April 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Achievement Tests; Comparative Analysis; Evaluation Methods; Higher Education; *Multiple Choice Tests; *Nurses; Test Construction; Testing; Testing Problems; *Test Reliability; *Test Validity

ABSTRACT

The purpose of this study was to compare certain characteristics of multiple-choice (MC) and complex multiple-choice (CMC) achievement tests designed to measure knowledge in medical-surgical nursing. Each of 268 junior and senior nursing students from four midwestern schools responded to one of four test forms. MC items were developed by converting original CMC items with four different systematic procedures. Results showed that: (1) students responded to five MC items for every four CMC items, (2) MC tests were at least as reliable as CMC tests though they did not measure exactly the same traits, and (3) CMC tests were at least as difficult as MC tests. Recommendations were made for test users. (Author)

COMPARATIVE RELIABILITIES AND VALIDITIES OF MULTIPLE CHOICE AND COMPLEX MULTIPLE CHOICE NURSING EDUCATION TESTS

Russell E. Dryden, Wichita State University

David A. Frisbie, University of Illinois

Presented at the Annual Meeting of the National
Council on Measurement in Education
Washington, D.C.
April 1975

COMPARATIVE RELIABILITIES AND VALIDITIES OF MULTIPLE CHOICE AND COMPLEX MULTIPLE CHOICE NURSING EDUCATION TESTS

The purposes of this study were to compare the reliabilities of multiple choice (MC) and complex multiple choice (CMC) achievement tests and to determine the concurrent validities of MC tests that were written to measure understandings of concepts and relationships in medical-surgical nursing. CMC items consist of a stem, a list of alternative responses called primary choices, and a list of responses called secondary choices, each of which is a combination of the primary choices. Students select their response for a CMC item from the list of secondary choices, only one of which is correct. The CMC format is illustrated in Figure 1 by Item 1A.

1A. Which of the following are frequent side effects of oral contraceptives?

a. Nausea
b. Dizziness
c. Headache
d. Weight gain
e. Breast discomfort

___1. a and b
___2. c and d
___3. All but e
_✓_4. All the above

1B. Which of the following are frequent side effects of oral contraceptives?

___1. Nausea and dizziness
___2. Headache and weight gain
___3. Dizziness and headache
_✓_4. All the above

Figure 1. Sample Complex Multiple-Choice Item Converted to Multiple-Choice Format

The major questions formulated as research hypotheses were:

1. Are MC and CMC achievement tests that were designed to measure the same objectives equally reliable?
2. What is the ratio of number of MC items attempted to the number of CMC items attempted by a group of examinees in a fixed period of time?
3. Is the correlation between individuals' MC and CMC subtest scores perfect (+1.00) when corrected for attenuation?
4. Are MC tests derived from CMC tests equally difficult?

Method

The CMC items used in this study were similar to those published to assist student nurses in reviewing for state licensure examinations and to provide guidance for nursing instructors in preparing classroom achievement tests. Sixty-four four-choice CMC items designed to measure knowledge, comprehension, and application in medical-surgical nursing were identified for test development purposes. The keyed secondary choice for a CMC item could consist of one, two, three, or all four primary choices. This relationship yielded four systematic procedures for converting CMC items to MC form. The 64 original CMC items were randomly split into two subtests, called C1 and C2 and were converted to MC subtests, called M1 and M2, respectively. Forms M1 and M2 were each comprised of eight items converted by each of the four procedures. The four final test forms, C1M2, C2M1, M1C2, and M2C1, contained 16 items of each of the four types and neither CMC or MC subtest consistently preceded the other.[1]

[1]Details of the item conversion procedures are in Dryden, 1974.

The subjects selected for testing were 212 junior and 56 senior nursing students at four midwestern schools of nursing. Three of the schools were hospital-affiliated and offered a diploma program. The fourth institution was an urban university with a baccalaureate degree program. Students were not randomly selected but all available students at these schools who were willing to participate were used. There is no reason to suspect that the group of subjects is vastly different from students in similar programs at other institutions.

Procedures

The study was designed to control various sources of random and systematic error. Each subject responded to only one test form and the four forms were randomly distributed in groups within each school. Subtest orders were counterbalanced. Explicit directions were read for each test administration and a stopwatch was used for timing the first 10 minutes of testing.

Subjects were stopped after 10 minutes of testing and were instructed to circle the number of the item they had been working on. Random marking of answer sheets was not observed and each subject was able to complete the examination.

Results

The ratio of the number of MC to CMC items that subjects attempted in the first 10 minutes of testing was determined to be 1.25. The median number attempted was 23.33 and 18.61 for MC and CMC, respectively.

Kuder-Richardson Formula 20 reliability coefficients computed for each of the eight subtests are reported in Table 1. The reliabilities of the MC subtests were adjusted with the Spearman-Brown Formula (n = 1.25) to equate

testing time. Each of the four adjusted MC reliability coefficients was larger than the corresponding CMC reliability coefficient. The difference were tested for statistical significance by computing 90 percent confidence intervals using a method developed by Feldt (1965). Table 2 is a display of the upper and lower bounds of the confidence intervals. In the two pairs of intervals which did not overlap, the MC reliability was higher than the CMC reliability in each case.

TABLE 1

$K\text{-}R_{20}$ Reliabilities for Final Subtest Forms

Test Form	Subtest		
	Complex Multiple-choice	Multiple-choice	
		Original	Adjusted
C1M2	.5991	.5692	.6228
M1C2	.3257	.3376	.3892
C2M1	.1878	.3328	.3840
M2C1	.3680	.5431	.5977

TABLE 2

Ninety Percent Confidence Intervals for $K\text{-}R_{20}$ Reliability Coefficients

Subtest	Test Form	Upper Limit	Lower Limit
C1	C1M2	.7037	.4828
M1	M1C2	.5486	.2121
C1	M2C1	.5329	.1829
M1	C2M1	.5448	.2054
C2	C2M1	.3998	-.0477
M2	M2C1	.7027	.4810*
C2	M1C2	.5017	.1302
M2	C1M2	.7212	.5134*

*Indicates the comparisons which did not overlap.

Since each subject received a MC and a CMC subtest score, a Pearson product-moment correlation was computed between subtest scores on each of the four test forms. Each correlation was adjusted for unreliability by correcting for attenuation. The original and corrected correlations are reported in Table 3.

TABLE 3

Correlation Coefficients for Multiple-choice and Complex Multiple-choice Subtest Scores on Each Final Test Form

Test Form	r_{mc}	$r^{a}_{\infty\,\infty}$	n
M1C2	.193	.582	67
M2C1	.423	.946	67
C1M2	.592	1.014	68
C2M1	.392	1.569	66

[a]Disattenuated correlation coefficients.

Ninety percent confidence intervals for the disattenuated coefficients were computed using a method developed by Forsyth and Feldt (1969). The upper and lower limits are given in Table 4. The hypothesis that the

TABLE 4

Ninety Percent Confidence Intervals for Disattenuated Correlation Coefficients

Test Form	$r_{\infty\,\infty}$	Est. Standard Error	Upper Limit	Lower Limit
M1C2	.582	.0490	1.0977	.9023
M2C1	.946	.0196	1.0397	.9603
C1M2	1.014	.0038	1.0117	.9883
C2M1	1.569	.1234	1.2461	.7539

disattenuated correlations do not differ from unity was not supported in any of the four cases.

A one-tailed t test was applied to test the differences in means on subtests which contained different but corresponding items. Means and standard deviations are shown in Table 5. The difference between the mean number correct on subtests M1 and C1 was not significant ($t = .401$, $df = 266$, $p > .05$). However, the difference between subtests M2 and C2 was significant ($t = 3.02$, $df = 266$, $p < .05$).

TABLE 5

Subtest Means and Standard Deviations

Test Forms	Subtest	Mean	Standard Deviation	N
C_1M_2 M_2C_1	C_1	16.15	3.46	135
C_2M_1 M_1C_2	M_1	16.31	3.03	133
C_2M_1 M_1C_2	C_2	16.31	2.90	133
C_1M_2 M_2C_1	M_2	17.56	3.77	135

Discussion

Conclusions drawn from the findings of this study should be regarded as tentative pending a replication of the study. The authors are not aware of other research reported regarding the questions studied here.

The results suggested that students can attempt five MC items in the time required to try four CMC items. In a 40-minute testing session, therefore, 93 MC or 74 CMC might be used if the relative responding rates of examinees are 5:4 beyond the first 10 minutes of testing. This would imply that a MC test is likely to better sample the content domain than is a CMC test when a given amount of testing time is available. The reliability evidence also indicated that the longer test is more reliable.

The fact that the MC and CMC reliabilities differed significantly in only two cases indicates that some factor other than item format was affecting the reliabilities. One factor that probably influenced the reliabilities of the original CMC subtests was the difficulty level of the items. The mean item difficulties (percent of the group responding incorrectly) on the four original CMC subtests were 48, 50, 51, and 48. These averages are too high for obtaining maximum reliability. If the item difficulties had averaged about 37.5, the items may have been higher in discrimination and, therefore, made for a more reliable test.

The MC-CMC subtest correlations were less than perfect. Though two of the disattenuated correlation coefficients were "close" for practical purposes, further research is needed before educational import can be attached to this finding. Though the converted items were similar to the original CMC items in content, they were not made up of corresponding converted items. There also was a problem with the reliabilities of the M1 and C2 subtests;

apparently the quality of the original CMC items was insufficient. Research on other item format comparisons (Frisbie 1973, Frisbie 1974) supports the notion that slightly different skills may be required of the examinee when item format varies. Further research is necessary before the extent of these differences and the specificity of the skills can be identified. The question of what is measured when a particular item format is employed certainly has a bearing on test validity in achievement testing situations.

Theoretically-derived chance scores on the MC and CMC tests used here were identical; subtest lengths and number of alternatives per item were the same. The conflicting results obtained when test difficulties were compared may have been produced by the relatively high item difficulties. Subjects could not answer many of the items correctly no matter which format the items were in. The findings regarding relative difficulties were at best inconclusive.

The results of this study suggest that more research in this area needs to be done if any sound conclusions are to be reached. A study comparing these two item formats, but using original CMC items of better quality than those used in this study, may yield more conclusive results. Factors present in the original items may have been the source of the difficulties in this study. Future studies might also include a valid external criterion in an attempt to clarify the validity question. If CMC and MC items do not measure the same skills and knowledge, which of the two is a better measure of the traits intended to be measured? Response rate with different item formats also merits further study. The data reported here reflect rate of response during the initial ten minutes of testing. The assumption has been made that this rate remains constant throughout the remainder of the testing period. The assumption actually represents an empirical question which should be addressed because it relates to projected test length and size of adjustment of the reliability estimate when testing time is held constant.

References

Dryden, R. E. A comparative study of multiple-choice and complex multiple-choice tests in nurses training. Unpublished master's thesis, Wichita State University, 1974.

Feldt, L. S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 30:357-369, 1965.

Forsyth, R. A. and Feldt, L. S. An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. *Educational and Psychological Measurement*, 29:61-71, 1969.

Frisbie, D. A. Multiple-choice versus true-false: A comparison of reliabilities and concurrent validities. *Journal of Educational Measurement*, 10:297-304, 1973.

Frisbie, D. A. The effect of item format on reliability and validity: A study of multiple-choice and true-false achievement tests. *Educational and Psychological Measurement*, 34:885-892, 1974.